

生成式人工智能犯罪的刑事责任分配探究

李华帅^{1,*}, 李文博¹

(1. 华东政法大学, 上海市, 201620; *通讯作者, 3194628524@qq.com)

摘 要:新一轮科技革命与产业变革曙光可见,在数字经济不断推进的大背景下,人工智能技术蒸蒸日上,并与多种应用场景深度融合,正成为推动人类进入智能时代的决定性力量。然而,人工智能的快速发展影响着刑法适用,生成式人工智能的刑事责任主体资格问题一直是学界争议的焦点。笔者认为,无论人工智能发展到何种程度,本质上仍是作为人类犯罪的工具,不具备刑事责任主体特征。对涉及人工智能犯罪做好相应预案,需要对生成式人工智能犯罪的刑事责任深度探究并正确分配以进行刑法规制。对于生成式人工智能犯罪,仍需坚持自然人主义,在责任分配的考量中,应当从三个核心层面出发,即设计环节的责任归属、生产过程的责任承担以及最终使用者的责任界定,从而确保责任的合理分摊和有效落实。

关键词:生成式人工智能犯罪;刑事责任;犯罪主体

引言

人工智能(Artificial Intelligence, AI),指的是通过计算机或受其控制的机器系统,来模拟、增强和拓展人类智能的一系列理论、方法、技术和应用。它涵盖了从感知环境、获取知识,到利用这些知识来达成最优结果的整个过程,是计算机科学的一个重要分支,致力于实现机器的智能行为。近年来,人工智能技术蓬勃发展,为新一轮科技革命强势赋能,正成为推动人类进入智能时代的决定性力量。“全球产业界充分认识到人工智能技术引领新一轮产业变革的重大意义,纷纷转型发展,抢滩布局人工智能创新生态[1]”,包括智慧诊疗、智慧法院、智能解答、元宇宙等,近期爆火的生成式人工智能(AI Generated Content, AIGC)也被快速投入到各种前端应用中。据《人工智能:科技与社会》预测:在未来的40年内,人工智能将超越人类智能水平,然而与之对应的是,人工智能的犯罪以其手段智能化而愈发严重。随着无人驾驶、ChatGPT等人工智能技术产品的更新迭代,如何对人工智能犯罪进行正确定位并以有效法律进行规制的理论问题日渐突出,人工智能犯罪的刑事责任分配问题也逐渐成为学者们研究的焦点。笔者认为,研究设计和使用人工智能的自然人无疑可以成为刑事责任的承担者,然而人工智能本身是否具有刑事责任主体的资格却值得深究。

目前,世界各国对于如何分配人工智能犯罪刑事责任的问题仍然有着较多争议。在2016年,欧盟司法委员会(JURI)颁布了《机器人领域的欧洲民法规则》(European Civil Law Rules in Robotics),并倡议设立一个专门的欧盟人工智能监管机构,以法律视角将更先进的自主机器人视作具有电子人格的法律实体。2017年,沙特阿拉伯授予了世界上首位“女性”机器人索菲亚(Sophia)“公民”身份,此举不仅引发了公众的热议,还激发了对人工智能在法律框架内享有有限电子人格及其对应权利的深入讨论。2021年,美国政府问责署(the U.S. Government Accountability Office, GAO)对联邦机构及其他组织机构发布了人工智能(Artificial Intelligence, AI)问责框架,反对人工智能以其作为机器的本质承担加于自然人的责任分配,关注程序正义以及针对损害的问责和追究。

人工智能技术的快速发展体现了加强人工智能问责领域研究的重要性和迫切性。当前我国对人工智能领域立法尚缺,学界对于刑事主体资格认定也未形成共识。尽管在当下,人工智能处于狭义阶段,“仅具有局限于特定领域的专用智能和感知智能,离以通用智能、认知智能为标志的广义人工智能特别是超级人工智能还有较大差距[2]”,人工智能系统还不是刑法规定的犯罪主体。然而,技术的发展速度是不可预料的,人工智能的不可控性具有带来刑事风险的隐患,需要应对未来可能出现的完全自主的广义人工智能系统是否作为犯罪主

体追究刑事责任进行理论研究。鉴于人工智能在法律框架内所扮演的角色日益凸显，深入探讨其在刑事法律体系中的主体资格问题，显得尤为关键。这一研究不仅有助于更好地理解人工智能的法律地位，还能为相关法规的制定和完善提供有力的理论支撑。

本文将力图厘清人工智能犯罪的主体性问题，论证人工智能并不具备刑事责任主体特征，并且论述涉人工智能犯罪时刑事责任的承担方式。

1 人工智能刑事责任主体地位的否定

关于人工智能是否具有主体性，学界通常有两个主流观点：肯定论和否定论。

1.1 肯定论

在展望未来的强人工智能时代，有一种观点认为高度智能化的人工智能将趋于接近甚至等同人类智能，因此主张应赋予其刑事法律主体的地位。这一“肯定论”被众多学者所倡导。

刘宪权教授便是其主要的理论支撑者。他的核心观点是：（一）强人工智能会在人类设定的程序和算法控制下产生自主学习能力，自发地认识世界；（二）与动物相比，强人工智能可以依照其自主的意志，选择性地实施危害社会的行为，其强大的数据分析能力能帮助其快速进行反应，具有辨认和控制能力，因此处罚强人工智能和处罚动物并不相同；（三）在强人工智能领域，可以设想一系列具有针对性的处罚措施，包括但不限于数据清除、程序调整乃至永久废止，以实现对违法行为的法律制裁。这些处罚方式将遵循法定程序，经过审判，并依据法律对其罪行进行定罪和量刑。由此可见，强人工智能在法律层面上具备成为刑事主体的资格[3]。

持肯定论观点的学者认为如下前提成立：

第一，从客观上看，智能机器人与自然人并没有遥不可及的距离。自然人与机器人的差别在于，自然人拥有情感、同理心和适应能力，可以自主学习并创造性的解决问题，而机器人则根据程序指令解决问题。然而，随着人工智能技术取得显著的进步，当它进入强人工智能的层次时，将展现出超越人类预设程序的自主性。在这一阶段，人工智能将不再完全受限于既定编程，而是能够基于其内在的、非外部操控的意愿，独立地、有选择性地执行行为。这一特性彰显了强人工智能的高度自主性和智能决策能力。因而具备了相应的辨认能力和控制能力。此时，智能机器人和自然人可能仅存在形体、机理上的区别[4]。

第二，“智能机器人没有自由意志”是一种主观臆断。自然人的自由意志是一种凭借高水平的思考、信念、价值观以及过去的经验来做出决定，并且对低水平的冲动加以控制的能力。因而自由意志具备两个特性：其一是不确定性，思考事物的方式会相应变化；其二是自主决定性，人的脑状态不会背离物理规律，有相应神经回路的函数值。通过程序、算法控制和大数据技术辅助下的数据获取技术，以及对自然人脑的反向仿制，不能排除未来机器人产生自由意志的可能性。

第三，自然科学的飞速进步提高了人工智能获得主体性地位的可能性。在技术的不断发展中，如果人工智能超越现有学习模式，拥有对自身意识系统反思的能力，理解自身意识系统的基本性质，就有可能对系统进行改造，创造新规则，可以逐步实现对人脑的模仿并产生人类智识，并赋予意志和道德，从而不得不考虑人工智能的人格问题。而由于自然环境被人类改造的舒适稳定，生物进化的步伐远不及自然科学发展的迅猛，人工智能在未来极有可能超越人类的智能和认知水平。基于这一前景，肯定强人工智能在法律上拥有主体性地位的趋势愈发明显。

第四，刑事责任主体不以自然人为唯一。换句话说，随着科技的飞速发展，不仅自然人，其他事物也展现出了作为刑事责任主体的潜在可能性。单位在刑事责任主体的认定上，经历了从“法人无犯罪能力”到“单位犯罪制度确立”的显著转变，其刑事责任能力的根源，实质上是人类意志的间接体现和延伸[5]。与此同时，作为现代科技产物的智能机器人，在执行人类预设程序的同时，甚至能进行算法之外的自主选择 and 决策，这无疑也是人类意志的一种高级延伸。值得注意的是，相较于单位与自然人的联系，人工智能与自然人意志的关联更为紧密和深远。尽管单位与自然人在本质上存在显著差异，但它依然能够独立承担刑事责任。以此类推，拥有高度自主性和决策能力的强人工智能，同样应当被视为具备成为刑事主体的潜在资格[6]。

综合上述观点，持肯定论的学者认为强人工智能具有主体性，可以对强人工智能进行问责。

1.2 否定论

否定论认为,在未来的世界,人工智能始终是人类设计的产物,不具有自主意识,不具有主体性。

否定论者主要分为“完全的否定论”者和“不完全的否定论”者。“不完全的否定论”者仅认为,弱人工智能不具备刑事责任主体性,而对强人工智能主体性持保留态度。这种观点并无参考价值,因为即使是“肯定论”者,也反对弱人工智能的刑事责任主体资格。“完全的否定论”者认为,所有人工智能都不具备刑事责任主体资格。这个观点与肯定论形成了鲜明的对立,更加有论证的必要,需要洞察人工智能特征以窥见其底层逻辑。

“完全的否定论”者观点主要如下:

其一、从当前的自然科学发展阶段来说,要成功研究出类似于人类的强人工智能还为时过早。目前人工智能存在诸多技术瓶颈与发展难题,由于现实的复杂性和人的主观能动性,强人工智能能否成功问世仍值得探讨,探讨强人工智能时代为时过早,把人工智能归为犯罪主体并无现实意义。

其二、对智能机器人赋予自主意志,只是肯定论者的主观臆想。肯定论者认为智能机器人具备与人脑类似的构造,从而进行深度学习和自由选择,继而可能产生人的意志。尽管刘宪权教授认为“绝对否定智能机器人具有自由意志的论断是非常武断和不可靠的[7]”,笔者认为,“绝对肯定智能机器人具有自由意志的论断是非常武断和不可靠的”,同样是把主观臆想上升为绝对真理。并且,人工智能实施了程序算法以外的行为,并不能必然的证明人工智能生成了自主意志,也可能是物理层面的机器故障。人工智能并不存在刑法意义上的自主意志,无法对自己行为的结果有情感与理智双层面的认知,不具有行为的目的性,缺乏相应的控制能力和辨认能力[8]。人工智能行为只是程序控制肢体的“机械行为”而为刑法所规制。

其三、将作为自然人集合体的单位与人工智能类比是引喻失义。肯定论者认为智能机器人与人类意志的联系性更强,人工智能的意志自由程度高于单位,无法否认人工智能的自主意志。社会承认纯粹拟制主体的刑事责任主体资格地位,也应当把人工智能作为犯罪主体加以归责。然而单位本身并无自由意志,单位由自然人组成,自然包含自然人的自由意志,可智能机器人并未包含自然人,何谈自由意志的存在?因此,以单位为类比,拟制人工智能的刑事主体的存在并不恰当[9]。

因此,否定论的学者认为人工智能无论发展到何种阶段,都只是人类的辅助工具,不具有自由意志。并且就目前发展阶段观之,惩罚强人工智能并无现实意义。刑法是由人制定的,刑法是为了人制定的,在探讨人工智能领域的发展时,应避免过度忧虑所谓的“奇点”现象,并审慎对待通过超前立法直接赋予人工智能刑事责任主体资格的做法[10]。超前立法不仅缺乏充分的法理依据,而且可能在实际操作中引发诸多争议和不确定性。应当依据科学和技术发展的实际情况,审慎、逐步地构建与人工智能相关的法律体系,确保法律的公正性、合理性和适应性。

1.3 人工智能犯罪主体地位之我见

在我国当前的刑法体系下,犯罪主体是指那些实际实施犯罪行为,并依法需要为其罪行承担相应责任的个体或单位。基于这一界定,犯罪主体需具备三大核心特征:其一,具有自我辨认和行为控制的能力;其二,客观上实施了危害社会的犯罪行为;其三,依据法律规定,应当承担刑事责任。此外,为了满足社会规制的需要,刑法可以运用拟制的方式赋予某种社会对象刑事主体地位。综上所述,笔者认为,人工智能独立承担刑事责任不具有理论自治性,理由如下:

1.3.1 对人工智能刑事责任能力的认定存疑

首先,人工智能不具有刑事责任能力。具有刑事责任能力是判断行为人是否承担刑事责任的前提,也是判断人工智能是否具有刑事主体性的关键条件。刑事责任能力取决于行为人对自已的行为及产生的后果是否具有辨认和控制能力。辨认能力是指行为人对自已行为的认知能力,包括从刑法规范评价的角度出发认识到自己行为的作用、性质以及造成的后果和影响。而控制能力是指行为人是否能够基于自主意志支配自身行为。笔者认为,虽然人工智能在形体上可以实行类似自主的行为,如通过系统发出语音、通过中央控制系统对周边环境施加影响等,但这其中的“意思”并不能归于人工智能的自由意志,更多的是属于按照人设计的程序和算法进行运作,由自然人进行主导控制,属于自然人的意思表示。人工智能和自然人可以进行情感、想象、知觉的

认知活动不同，只能通过模型进行运算分析活动，而且不能对伦理的非难进行回应，不具有作为责任必要条件的自我反省能力，因而和自然人的认知能力存在巨大差异。尽管人工智能有远超自然人的数据收集能力和数据分析能力，但是“其本身对刑法规范既无理性的认识与遵从能力，也不能体会到违反规范所造成否定性评价带来的消极后果[5]”，因此，人工智能不具有刑事责任能力。

其次，无法对人工智能的刑事责任能力进行认定。无论是传统的四要件理论或是德日的阶层论，对于行为人是否具有刑事责任能力都同时需要考虑其刑事责任年龄和精神状态。目前我国刑法规定的刑事责任年龄是以自然人为中心的，而没有对人工智能刑事责任年龄的计算标准。对精神状态的分析需要结合生物学与心理学的知识，因此同样无法判断人工智能的精神状态。

综上所述，笔者对人工智能的刑事责任能力持否定性态度，因此人工智能不具有刑事责任主体资格。

1.3.2 人工智能缺乏刑罚能力

人工智能不会受到刑罚的威慑，无法实现刑罚目的。刑罚的目的是惩罚和预防犯罪。

首先，对于人工智能的惩罚目的并不具有可行性。针对人工智能的罚金刑，专家提出一种观点，即通过要求人工智能的设计者、制造者以及使用者采取如强制购买保险等措施来履行法律责任。然而，这一方法实际上是将对人工智能的处罚责任转移至其他个体之上，此举与刑法的责任自负原则相悖。至于人工智能的自由刑，笔者认为，对人工智能实施自由刑并不能达到与人类相同的惩戒效果，因为人工智能本身并不具备理解自由含义以及刑罚意图的能力。其运行完全基于预设的算法和程序，无法像人类那样感受到自由的价值和刑罚的威慑力；关于人工智能的死刑，人工智能以是否存在作为生命的判断标准，没有可以施加死刑的基础。

其次，需知对人工智能的惩罚并不能实现预防犯罪的根本目的。在我国，对自然人犯罪的处罚已构建了一整套成熟的刑罚体系。旨在通过一般预防和特殊预防来遏制犯罪。一般预防旨在教育并警示那些可能步入犯罪道路的社会不稳定分子；特殊预防则是通过对犯罪者的惩罚，使其产生畏惧和悔罪心理，防止其再次犯罪。然而，人工智能的本质属性与自然人截然不同。它没有生命，没有独立的财产，更不具备独立的意识。它是由各种零件和技术构成，其运行完全基于预设的代码算法和程序。因此，对人工智能实施刑罚不仅无法使其体验到刑罚的痛苦，更无法激发其悔罪或畏惧心理。同时，这样的惩罚也无法对其他潜在的犯罪者产生威慑作用，从而削弱了民众对法律的信赖，并降低了刑罚的严肃性[11]。综上所述，从刑罚目的的角度出发，人工智能无法被认定为犯罪主体。

1.3.3 人工智能欠缺拟制条件

一些学者提出，把法人作为犯罪主体的逻辑框架可以为人工智能的发展提供借鉴[6]。笔者认为，由于人工智能特定条件的限制，无法完全沿用拟制法人主体的理论路径。

首先，法人具有独立的利益诉求，这是其成为犯罪主体的前提之一。单位犯罪的成立，往往基于犯罪人为了单位的利益而谋求非法所得，或犯罪所得主要归单位所有。在此意义上，法人拥有确定的财产权；然而，人工智能并不具备独立的财产权利，它更多地是作为财产的一部分存在，而非独立的利益主体。

其次，法人还拥有独立自主的意志。尽管法人由自然人组成，但在决策和行为时，它是以法人的名义，按照规定的程序，通过决策机构的讨论和决定来体现其整体意志。这种意志虽然源于自然人，但又不完全等同于自然人的意志。然而，人工智能的行为是基于预设的算法和程序，它缺乏自主决策和意志的能力，其运行方式也并未经过类似法人决策机构的协调和统一。

2 人工智能犯罪的刑事责任分配

基于上文的分析，人工智能并不具备成为刑事主体的必要条件。在对人工智能犯罪进行刑法规制时，应当以传统刑法理论作为基础，坚持自然人主义，坚持刑法的谦抑原则，清晰界定并明确相关人员的刑事责任，以确保责任归属的准确性和公正性。

2.1 人工智能设计者的责任

在探讨人工智能犯罪时，人工智能设计者的行为虽非最直接原因，但实为犯罪的初始诱因。如若设计者出于实现特定犯罪目的而设计人工智能，并为其编制了实施犯罪的系列程序，且输入大量犯罪相关数据，即便明知其产品可能带来社会危害，仍抱有放任或期望其发生的态度，那么，此类人工智能犯罪实则是设计者意图和意志的延伸，由设计者承担故意犯罪的责任自属合理。除此以外，若设计者在设计过程中未能尽到应有的审慎的注意义务，或虽意识到人工智能可能产生危害行为，却过分自信地认为能够避免，则应承担过失的刑事责任。不论是设计过程中存在的固有缺陷，还是人工智能在深度学习过程中偏离预定程序的行为，均反映出设计者在履行注意义务上的过失。

2.2 人工智能生产者的责任

对生产者的刑事责任承担，也应当分为故意犯罪和过失犯罪。首先，若生产者在制造过程中明知其人工智能产品可能产生危害社会的后果，却执意制造，此种行为应被判定为故意的刑事责任。具体而言，生产者的这一生产行为，实质上可视为对设计者设计的人工智能产品犯罪的帮助行为，从而应被界定为帮助犯。其次，从另一个角度审视，若生产者在制造过程中未能充分履行其应有的安全保障义务和告知义务，导致了不良后果，那么生产者应承担过失的刑事责任。这种过失体现在未能有效预防和避免潜在的社会危害。目前针对人工智能的注意义务主要是避免义务和监督义务。生产者有妥善管理产品的义务，在生产销售及售后阶段，都需要保障人工智能产品处于可控有益的状态，并且要跟进持续检测。

2.3 人工智能使用者的责任

在探讨人工智能使用者的刑事责任时，其责任界定依赖于不同的使用行为方式。首先，如果使用者出于故意，把人工智能产品作为其实施犯罪的手段或工具，那么，无疑应当承担故意的刑事责任。其次，由于使用者的过失导致了人工智能的犯罪行为，例如疏忽大意疏于监管等，使用者就需要承担过失的刑事责任。

然而，使用者的注意义务应当有别于设计者和生产者。设计者和生产者具备人工智能专业领域相当的专业知识，因而所具有的注意义务要高于作为社会一般人的使用者。如果设计者和生产者详细告知了使用者注意事项，履行了提醒告知的义务，便不影响犯罪后使用者刑事责任的承担。反之，使用者对相对注意的事项毫不知情，那么使用者不具有对人工智能犯罪的预见可能性，则不能承担刑事责任。

2.4 刑罚处罚方式的适当扩张

前文说到，人工智能不具备独立的刑事主体资格，因此这里提到的刑罚处罚方式的适当扩张仍然是针对自然人主体保护受害者利益的角度提出的。

首先，对人工智能的设计者进行职业限制。根据人工智能产品的社会危害性大小，决定一定时间维度内对其禁止从事人工智能领域的设计工作。其次，前文提到惩罚无法让人工智能感受到痛苦，但可以要求删除修改变更其运行程序或者对其彻底销毁，本质上这些惩罚还是回到了自然人身上，遭受了财产内容的损失。

3 结 语

大风泱泱，大潮滂滂。人工智能的快速发展是时代所趋，是时代所向。人工智能固然以其强大的内驱力推动着社会生产力的发展，然而对其隐藏的风险要保持警惕和审慎的态度。如果人工智能被随意滥用，或者成为犯罪分子用于犯罪的利器，将极大地破坏社会秩序。正确处理人工智能犯罪的刑事责任分配问题，是当下亟需解决的时代命题。无论人工智能发展到何种阶段，其本质上都只是人类犯罪的工具，不具有刑事主体资格，在针对人工智能的刑事处罚问题上，应当坚持自然人主义的原则，确保相关责任能够由实际涉及的人员来承担。笔者提出了几点要点：一是质疑人工智能的刑事主体地位，明确其非主体性；二是明确在人工智能造成损害时，应如何正确分配刑事责任；三是探讨刑罚处罚方式的合理扩展，以适应新型犯罪形态。这些要点旨在明确归责机制，对人工智能犯罪的刑法规制提供有益的参考。

参考文献

- [1] 谭铁牛. 人工智能的历史、现状和未来[EB/OL]. (2019-02-18) [2024-05-19]. https://www.cas.cn/zjs/201902/t20190218_4679625.shtml.
- [2] 皮勇. 人工智能刑事法治的基本问题[J]. 比较法研究, 2018(5): 149-166.
- [3] 刘宪权. 人工智能时代的刑事风险与刑法应对[J]. 中国检察官, 2018(9): 79.
- [4] 刘宪权, 房慧颖. 涉人工智能犯罪的前瞻性刑法思考[J]. 安徽大学学报(哲学社会科学版), 2019, 43(1): 108-117.
- [5] 张明楷. 刑法学[M]. 北京: 法律出版社, 2024: 101.
- [6] 刘洪华. 人工智能法律主体资格的否定及其法律规制构想[J]. 北方法学, 2019, 13(4): 56-66.
- [7] 刘宪权. 对强智能机器人刑事责任主体地位否定说的回应[J]. 法学评论, 2019, 37(5): 113-121.
- [8] 陈洪兵. 人工智能刑事主体地位的否定及实践展开——兼评“反智化批判”与“伪智化批判”之争[EB/OL]. (2023-10-14)[2024-05-19]. <https://zhuanlan.zhihu.com/p/661260824>.
- [9] 时方. 人工智能刑事主体地位之否定[J]. 法律科学(西北政法大学学报), 2018, 36(6): 67-75.
- [10] 吴波, 俞小海. 人工智能时代刑事责任认定思路的挑战与更新[J]. 上海政法学院学报(法治论丛), 2018, 33(5): 95-101.
- [11] 王文明, 齐卫红. 论人工智能的刑事主体地位与刑法应对[J]. 河南工业大学学报(社会科学版), 2021, 37(5): 52-58.